Navigating challenges and solutions of communicative eXplainable AI

Yizhe Huang
School of Intelligence Science and Technology
Peking University
szhyz@pku.edu.cn

Abstract

This essay presents a framework for modeling eXplainable AI (XAI) as a communication process, outlining associated challenges such as explainability of explanations and gaining human trust. Furthermore, it offers potential solutions to address these challenges. The essay posits that endowing XAI systems with cognitive abilities and structured representations holds promise in enhancing their efficacy.

1 Introduction

As AI technology continues to permeate every facet of human society, our reliance on AI-generated content grows, prompting a heightened focus on the explainability of AI output. First, there is a drive to uncover the scientific underpinnings of AI's conclusions. Understanding the process through which AI arrives at its decisions is crucial for humans as they seek to assimilate this knowledge. The transfer of such experience and knowledge is vital for the ongoing progress and survival of human society. Second, humans need AI to provide certain explanations for their actions so that humans can trust AI's decisions. For instance, in scenarios such as large financial transactions facilitated by AI, it is imperative for individuals to have insight into the rationale behind AI-generated recommendations in order to place full confidence in its decisions. With AI's increasing capabilities and integration into society, AI researchers are already worried about the security issues that may be caused by artificial intelligence [6]. XAI plays a pivotal role in augmenting human comprehension and oversight of AI decision-making processes, and its significance is set to escalate.

One of the classic approaches to XAI involves modeling it as a communication process [7, 3]. This dynamic entails the exchange of information between an explainer (which is often a machine in XAI) and an explainee (which could be a human or another automatic system). Throughout this process, the explainee can seek specific information from the explainer to enhance their understanding, while the explainer endeavors to furnish comprehensive explanations. It is important to note that at times, the explainer may offer biased or incorrect information in an attempt to gain the explainee's trust.

A framework proposed in [3] (refer to Fig. 1) divides the explainer's explanation process into cognitive and social processes. The former determines the requisite information for the explanation, encompassing factors such as the environment, one's mental state, and the mental states of others. The latter involves providing the explanation to the explainee, drawing upon the information assimilated during the cognitive process. Subsequently, the explainee seeks clarification from the explainer based on the information received, as well as any perceived contradictions or uncertainties. [3] introduces varying levels of explanation, spanning from observation data (0-level) to one's intentions (1-level), beliefs regarding one's and others' mental states (2-level), and broader social group characteristics (N-level). Additionally, meta-explanation, which denotes explanations for explanations, is explicitly highlighted as a key component.

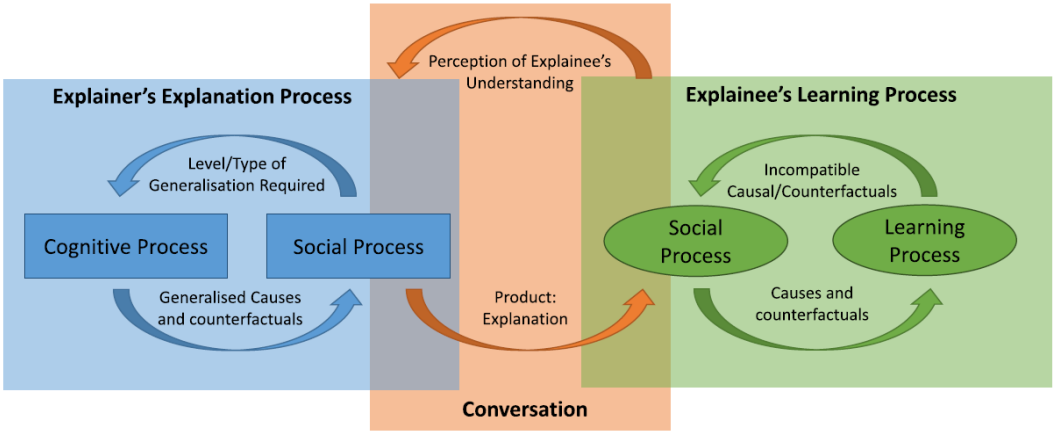

Figure 1: Explanation as a conversational process. It is Figure 2 of [3].

2 Some Challenges in Communicative XAI

2.1 The explainability of explanations

The critical issue in XAI lies in the mechanism of generating explanations, particularly the meta-explanation process. While existing models can provide reasonable explanations for their outputs, such as using the "Think step by step" prompt technique [5] for mathematical problems, there are concerns about the reliability of these explanations. For instance, it has been observed that large language models (LLMs) may not consistently provide accurate derivation processes, even if the answers they yield are correct[1]. This casts doubt on the reliability of the explanation generation mechanism, suggesting that it may simply imitate the explanation rather than authentically elucidating the process of arriving at the answer.

Moreover, the lack of transparency in the explanation generation mechanism raises security concerns. When the generation mechanism of an explanation is itself unexplainable, it becomes challenging to trust that the explanation has not been manipulated to deceive users into accepting a harmful answer.

To ensure that explanations genuinely correspond to the process of arriving at an answer, it is essential that the mechanism for obtaining the answer possesses a meaningful structure. This could involve representations such as logical formulas, programming languages, or And-or graphs. Even within neural networks, introducing structured controls over intermediate variables, such as learning a set of discretized features in VQ-VAE or leveraging inherent graph structures in graph neural networks, can enhance explainability. Alternatively, providing the module responsible for explanation with a certain structure, such as techniques to discern the meaning of intermediate variables in black-box neural networks or structurally decomposing the answer itself, can also contribute to improving explainability.

2.2 Gain the trust of humans

Gaining the trust of human users is one of the most important goals, if not the most important, of XAI. This trust operates at two levels: trust in the capabilities of AI and trust in the values it upholds. As AI technology advances, there is a growing belief in its capabilities, with a rising concern regarding the values embedded within AI systems. This concern has propelled the topic of value alignment to the forefront of discussions in recent years [10, 4].

Value alignment within Communicative XAI encompasses the cognitive process illustrated in Figure 1, principally focusing on aligning and interpreting mental states. This necessitates the presence of a psychological system within the XAI framework, whether explicitly or implicitly expressed. Moreover, the XAI system must possess Theory of Mind (ToM) capabilities, an area that is still underdeveloped in AI. Efforts in XAI have explored employing Bayesian inference [9] or neural networks [1] as ToM modules, albeit with limitations in their applicability and a prerequisite of task-specific prior knowledge.

[1] See my essay for the "Communication" lecture for details.

An advanced capability within this realm is active alignment, where the explainer is not always the party who is passively questioned, but actively engages in reducing uncertainty through exploratory behaviors or inquiries, leading to an increase in common ground. This bidirectional alignment framework holds promise for fostering mutual understanding [10].

An essential aspect that warrants attention is the measurement of human trust and the explainability of AI models. These properties bear subjective elements, necessitating measurement through human studies. While some methods directly inquire about human trust levels, the outcomes are notably influenced by the subject population's distribution. An alternative, more objective approach involves investigating human mental states and correlating them with AI-inferred results. Moreover, certain articles have underscored the limitations of subjective measures [2] and have endeavored to propose entirely objective metrics [8].

3 Discussion

We review a framework depicting XAI as a communication process and discuss its associated challenges. Primarily, the XAI system necessitates specific cognitive capabilities to align with human cognition and psychological states. While structured representations hold promise for XAI, their scalability emerges as a prominent challenge. Consequently, striking a balance between explainability and scalability becomes imperative, prompting the need to make trade-offs in this regard.

References

[1] Arjun R Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience*, 25(1), 2022. 2

[2] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020. 3

[3] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, 2021. 1, 2

[4] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Heng Lou, Kaile Wang, Yawen Duan, Zhongshi He, Jianfeng Zhou, Zhaowei Zhang, Fangfang Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yaguo Lei, Hongyao Xu, Brian Tse, Jie Fu, Stephen McAleer, Yanfei Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey. *arXiv (Cornell University)*, October 2023. doi: 10.48550/arxiv.2310.19852. URL `https://arxiv.org/abs/2310.19852`. 2

[5] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2

[6] Future of Life Institute. Pause giant ai experiments: An open letter - future of life institute, November 2023. URL `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`. 1

[7] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. 1

[8] Francesco Sovrano and Fabio Vitali. An objective metric for explainable ai: how and why to estimate the degree of explainability. *Knowledge-Based Systems*, 278:110866, 2023. 3

[9] Samuel Westby and Christoph Riedl. Collective intelligence in human-ai teams: A bayesian theory of mind approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6119–6127, 2023. 2

[10] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183, 2022. 2, 3